

**WG2 N4792R2  
L2/17-076R2**

***Date: 2017-05-9***

**Title: Revised proposal for the encoding of an Egyptological YOD and Ugaritic characters**

**Source: Michel Suignard**

**Status: Individual Contribution**

**Distribution: UTC, WG2**

**References: WG2 N2043 (L2/99-246), N2241(L2/00-293), N3382R(L2/08-049R),  
N3431(L2/08-087), N3432(L2/08-146), N3487(L2/08-272)**

**Summary:** This document contains a proposal for the encoding of an Egyptological YOD and 2 additional characters (all three in small and capital letter) in the Latin Extended-D block. While these characters have been proposed before and rejected, the proposed alternative solutions have not been implemented by the vendors, therefore harming the user community and it is time to come back to an easier solution which is the encoding of non-decomposable characters as originally proposed in N3487.

## **Introduction**

The story of the Egyptian transliteration characters goes a long way back. These characters were first introduced in document WG2 N2043 and finally proposed for encoding in document N2241, almost 17 years ago. Four of the six proposed characters were encoded in the Latin Extended-D block, but the proposed Egyptological YOD was not accepted because it was thought that it could be represented by a combining sequence using the Latin Letter I followed by either U+0313 COMBINING COMMA ABOVE or U+0357 COMBINING RIGHT HALF RING ABOVE. None of them fully represented the visual aspect of the character. Then it was suggested to use U+0486 COMBINING CYRILLIC PSILI PNEUMATA as the combining diacritic for this purpose, making its script property 'Inherited' (instead of 'Cyrillic'), and adding an annotation to the code point in the name list (ref N3382R). These suggestions were consequently rejected by document N3431, endorsed by N3432, recommending '*a specialized font for Egyptian transliteration be created that would permit the appropriate placement of the diacritic*'.

Then a new document N3487 was created requesting the additions of 6 characters (LATIN LETTER A, I, and U with SPIRITUS LENIS in both capital and small forms), with the LATIN SMALL LETTER I WITH SPIRITUS LENIS annotated: "used for Egyptological yod". The whole set was to be used for transliteration of the Ugaritic writing system. This was not accepted by the UTC, but instead, the previously rejected script property change for U+0486 COMBINING CYRILLIC PSILI PNEUMATA to 'Inherited' was accepted (but no added annotation).

The forth and back has apparently confused implementers because there is no commonly accepted representation of these letters (nothing worse than having an undefined representation model for the diacritic mark).

Making the matter worse is the fact most Egyptian transliteration are done in an italicized style, and none of the commonly available commercial fonts produce a result which is acceptable when combining an Italic 'i' with any of the suggested diacritics. Clearly the Cyrillic PSILI PNEUMATA has the right shape in term of diacritic, but the confusion brought by documents N3431, 3432, and the reversal done later by the UTC has ensured that there is no commercial solution. Furthermore, a specialized font for Egyptian transliteration with no clear guidance for the encoding of the sequence has no chance to be used in an interoperable fashion.

The author faced the issue when creating an extension of the Egyptian Hieroglyphs and trying to transliterate the ideograms and logograms associated with the proposed characters and add them in a database that can be shared with the user community. There was simply no good solution. There are fonts using either incorrect code points or PUA for the YOD. Alternatively, many users just go around and use the letter ‘j’. Two alternative solutions have been considered for the need of the YOD representation and the Ugaritic transliteration:

**1) Encode all 6 characters originally proposed by N3487 (suggested code points are not available anymore)**

À	A790	LATIN CAPITAL LETTER A WITH SPIRITUS LENIS
à	A791	LATIN SMALL LETTER A WITH SPIRITUS LENIS
		• used in transliteration of Ugaritic
Ĭ	A792	LATIN CAPITAL LETTER I WITH SPIRITUS LENIS
ĭ	A793	LATIN SMALL LETTER I WITH SPIRITUS LENIS
		• used for Egyptological yod
Ū	A794	LATIN CAPITAL LETTER U WITH SPIRITUS LENIS
ū	A795	LATIN SMALL LETTER U WITH SPIRITUS LENIS
		• used in transliteration of Ugaritic

With the understanding that LATIN SMALL LETTER I WITH SPIRITUS LENIS would be the Egyptian YOD. The text ‘SPIRITUS LENIS’ is a rarely used Latin-based term for ‘smooth breathing’, which would be ‘Psili pneumata’ in Greek. Obviously PSILI PNEUMATA’ could never be used in the name of a non-decomposable character. Rationale for these characters is provided in N3487 and is not repeated here.

It should be noted that the common representations for the YOD look more like *Ĭ* and *ĭ* because Egyptian transliteration is commonly italicized. This fact could be mentioned in an annotation in the name list because the code chart glyphs would de facto be extremely rare. In addition, the mark is always on the left of the I in capital form, unlike what would be typically done for a mark added to a capital I (on top).

**2) Encode a single COMBINING SPIRITUS LENIS.**

That solution could introduce a risk of some confusion with the already encoded U+0486 COMBINING CYRILLIC PSILI PNEUMATA, but the main point of adding a new diacritical mark would be to provide strict guidance concerning the rendering of the ‘smooth’ breathing in conjunction with the 3 vowels A, I, and U, including the italicized form of the small letter ‘i’ used for Egyptological YOD.

After UTC discussion, the preferred solution is to encode 6 new code points.

The code positions could be U+A7BA and U+A7BB (next available code positions), properties as follow:

```
A7BA;LATIN CAPITAL LETTER A WITH SPIRITUS LENIS;Lu;0;L;;;;N;;;A7BB;
A7BB;LATIN SMALL LETTER A WITH SPIRITUS LENIS;Ll;0;L;;;;N;;;A7BA;;A7BA
A7BC;LATIN CAPITAL LETTER I WITH SPIRITUS LENIS;Lu;0;L;;;;N;;;A7BD;
A7BD;LATIN SMALL LETTER I WITH SPIRITUS LENIS;Ll;0;L;;;;N;;;A7BC;;A7BC
```

A7BE;LATIN CAPITAL LETTER U WITH SPIRITUS LENIS;Lu;0;L;;;N;;;A7BF;  
A7BF;LATIN SMALL LETTER U WITH SPIRITUS LENIS;Ll;0;L;;;N;;;A7BE;;A7BE

They are not confusable with any existing single code points, but the new characters could be confused with the sequence of the corresponding Latin, Greek, or Cyrillic letters followed by a diacritical mark resembling the breathing mark. Given the existing confused situation in rendering these sequences, this does not exacerbate the risk in a meaningful way. There is a limited need to use a YOD or the Ugaritic transliteration characters in general purpose identifiers; their main purpose is to represent transliteration of historic Egyptian and Ugaritic text originally written using their native glyphs.

The script property for U+0486 COMBINING CYRILLIC PSILI PNEUMATA can keep the value 'Inherited', also it is likely to only be used with Cyrillic characters.

In the Egyptian transliterated alphabet, the Egyptological yod is sorted after the Egyptological aleph (U+A723), and behaves like a normal letter. But default ordering can be done following principle for newly added Latin characters. These proposed characters would have no decompositions into a base letter followed by a diacritic.

For text segmentation, and related properties these characters should be processed as regular Latin characters.

**Acknowledgement:**

This proposal is built on research done by Michael Everson and Bob Richmond which have contributed all the previous documents and investigated the need for these characters. This proposal could not have been done without their work for all these years.

## A. Administrative

1. Title:	<i>Proposal to encode Egyptological Yod and similar characters in the UCS</i>
2. Requester's name:	<i>Michel Suignard</i>
3. Requester type (Member body/Liaison/Individual contribution):	
4. Submission date:	<i>2017-05-09</i>
5. Requester's reference (if applicable):	
6. Choose one of the following:	
This is a complete proposal:	<i>Yes</i>
(or) More information will be provided later:	<i>No</i>

## B. Technical – General

1. Choose one of the following:	
a. This proposal is for a new script (set of characters):	<i>No</i>
Proposed name of script:	
b. The proposal is for addition of character(s) to an existing block:	<i>Yes</i>
Name of the existing block:	<i>Latin Extended-D</i>
2. Number of characters in proposal:	<i>6</i>
3. Proposed category (select one from below - see section 2.2 of P&P document):	
A-Contemporary <input checked="" type="checkbox"/> <i>A</i>	B.1-Specialized (small collection) <input type="checkbox"/>
C-Major extinct <input type="checkbox"/>	D-Attested extinct <input type="checkbox"/>
E-Minor extinct <input type="checkbox"/>	F-Archaic Hieroglyphic or Ideographic <input type="checkbox"/>
G-Obscure or questionable usage symbols <input type="checkbox"/>	B.2-Specialized (large collection) <input type="checkbox"/>
4. Is a repertoire including character names provided?	<i>Yes</i>
a. If YES, are the names in accordance with the “character naming guidelines” in Annex L of P&P document?	<i>Yes</i>
b. Are the character shapes attached in a legible form suitable for review?	<i>Yes</i>
5. Fonts related:	
a. Who will provide the appropriate computerized font to the Project Editor of 10646 for publishing the standard?	<i>Michael Everson</i>
b. Identify the party granting a license for use of the font by the editors (include address, e-mail, ftp-site, etc.):	<i>Michael Everson</i>
6. References:	
a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided?	<i>Yes, N3487</i>
b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached?	<i>Yes, by reference</i>
7. Special encoding issues:	
Does the proposal address other aspects of character data processing (if applicable) such as input, presentation, sorting, searching, indexing, transliteration etc. (if yes please enclose information)?	<i>yes</i>

## 8. Additional Information:

Submitters are invited to provide any additional information about Properties of the proposed Character(s) or Script that will assist in correct understanding of and correct linguistic processing of the proposed character(s) or script. Examples of such properties are: Casing information, Numeric information, Currency information, Display behaviour information such as line breaks, widths etc., Combining behaviour, Spacing behaviour, Directional behaviour, Default Collation behaviour, relevance in Mark Up contexts, Compatibility equivalence and other Unicode normalization related information. See the Unicode standard at <http://www.unicode.org> for such information on other scripts. Also see Unicode Character Database ( <http://www.unicode.org/reports/tr44/> ) and associated Unicode Technical Reports for information needed for consideration by the Unicode Technical Committee for inclusion in the Unicode Standard.

**C. Technical - Justification**

1. Has this proposal for addition of character(s) been submitted before? If YES explain	<i>N3487, but rejected for alternative solution</i>	<i>Yes</i>
2. Has contact been made to members of the user community (for example: National Body, user groups of the script or characters, other experts, etc.)? If YES, with whom? If YES, available relevant documents:	<i>Ireland NB, Egyptologists</i> <i>N3487</i>	<i>Yes</i>
3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included? Reference:	<i>Egyptologists, Copticists, Semiticists, and other scholars</i>	
4. The context of use for the proposed characters (type of use; common or rare) Reference:	<i>Used historically and in modern editions.</i>	
5. Are the proposed characters in current use by the user community? If YES, where? Reference:	<i>Yes</i>	
6. After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP? If YES, is a rationale provided? If YES, reference:	<i>Accordance with the roadmap. Keep with other Egyptological characters</i>	<i>Yes</i> <i>Yes</i>
7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)?		<i>No</i>
8. Can any of the proposed characters be considered a presentation form of an existing character or character sequence? If YES, is a rationale for its inclusion provided? If YES, reference:	<i>No</i>	<i>No</i>
9. Can any of the proposed characters be encoded using a composed character sequence of either existing characters or other proposed characters? If YES, is a rationale for its inclusion provided? If YES, reference:		<i>No</i>
10. Can any of the proposed character(s) be considered to be similar (in appearance or function) to, or could be confused with, an existing character? If YES, is a rationale for its inclusion provided? If YES, reference:		<i>No</i>
11. Does the proposal include use of combining characters and/or use of composite sequences? If YES, is a rationale for such use provided? If YES, reference: Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided? If YES, reference:		<i>No</i>
12. Does the proposal contain characters with any special properties such as control function or similar semantics? If YES, describe in detail (include attachment if necessary)		<i>No</i>
13. Does the proposal contain any Ideographic compatibility characters? If YES, are the equivalent corresponding unified ideographic characters identified? If YES, reference:		<i>No</i>

----